Семинар 2

Задача 1

T1 Дана выборка $x_1, \ldots x_n$ из неизвестного абсолютного непрерывного распределения. Для оценивания плотности в точке 0 используется оценка

$$\hat{p}_n(0) = \frac{\sharp\{i : x_i \in (-h/2, h/2]\}}{nh}.$$

(i) Докажите, что

$$MSE(\hat{p}_n(0)) := \mathbb{E}\Big[\big(\hat{p}_n(0) - p(0)\big)^2\Big]$$
$$= \left(\frac{(p''(0))^2}{576}h^4 + \frac{p(0)}{nh}\right)(1 + o(1)), \quad n \to \infty,\ h \to 0.$$

Запишем формулу bias-variance tradeoff:

$$MSE(\hat{p}_n(0)) = \mathbb{E}[(\hat{p}_n(0) - p(0))^2] = Var(\hat{p}_n(0)) + Bias^2(\hat{p}_n(0))$$

$$\hat{p}_n(0) = \frac{\#\{i : x_i \in (-h/2, h/2]\}}{nh} = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}\{-h/2 < x_i \le h/2\}$$

$$\xi_i := \mathbb{1}\{-h/2 < x_i \le h/2\} = \begin{cases} 1, & \mathbb{P}\{-h/2 < x_i \le h/2\} \\ 0, & 1 - \mathbb{P}\{\ldots\} \end{cases} \Rightarrow \xi_1, \xi_2, \ldots \xi_n \sim \text{iid } \mathcal{B}(1, \overbrace{\mathbb{P}\{-h/2 < x_1 \le h/2\}}^{p_h})$$

$$\Rightarrow \sum_{i=1}^{n} \mathbb{1}\{-h/2 < x_i \le h/2\} \sim \mathcal{B}(n, p_h)$$

- $$\mathbb{E}[\hat{p}_n(0)] = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n}\mathbb{1}\{-h/2 < x_i \le h/2\}\right] = \frac{1}{nh}\mathbb{E}\left[\sum_{i=1}^{n}\xi_i\right] \overset{\text{iid}}{=} \frac{1}{\cancel{n}h}\cdot\cancel{n}\,\mathbb{E}[\xi_1] = \frac{1}{h}\mathbb{P}\{-h/2 < x_i \le h/2\}$$

$$p_h = \mathbb{P}\{-h/2 < x_1 \le h/2\} = \int_{-h/2}^{h/2} p(x)\,dx = \int_{-h/2}^{h/2}\Big(p(0) + p'(0)x + \frac{1}{2}p''(0)x^2 + \overbrace{\bar{\bar{o}}(x^2)}^{x^2\cdot\bar{\bar{o}}(1)}\Big)dx$$

$$= p(0)\cdot h + 0 + \frac{1}{2}p''(0)\cdot\frac{1}{3}\left(\frac{h^3}{8} + \frac{h^3}{8}\right) + \bar{\bar{o}}(h^3)$$

$$= p(0)\,h + p''(0)\cdot\frac{h^3}{24} + \bar{\bar{o}}(h^3)$$

$$\mathbb{E}[\hat{p}_n(0)] = \frac{1}{h}\cdot\left(p(0)h + p''(0)\cdot\frac{h^3}{24} + \bar{\bar{o}}(h^3)\right) = p(0) + p''(0)\cdot\frac{h^2}{24} + \bar{\bar{o}}(h^2)$$

$$\Rightarrow Bias^2(\hat{p}_n(0)) = \left(p''(0)\cdot\frac{h^2}{24} + \bar{\bar{o}}(h^2)\right)^2 = \frac{h^4(p''(0))^2}{24^2}(1 + \bar{\bar{o}}(1))$$

$$\bullet\ \mathrm{Var}(\hat p_n(0)) = \mathrm{Var}\Big(\frac{1}{nh}\sum_{i=1}^{n}\xi_i\Big) \overset{iid}{=} \frac{1}{n^2h^2}\cdot n\,\mathrm{Var}(\xi_1) = \frac{1}{nh^2}\,p_h(1-p_h)$$

$$= \frac{1}{nh^2}\Big(p(0)h + p''(0)\frac{h^3}{24} + \bar{\bar o}(h^3)\Big)\Big(1 - p(0)h - p''(0)\frac{h^3}{24} - \bar{\bar o}(h^3)\Big)$$

$$= \frac{1}{n}\Big(\frac{p(0)}{h} + \underbrace{p''(0)\cdot\frac{h}{24}}_{\to 0,\ h\to 0} + \underbrace{h\bar{\bar o}(1)}_{\to 0}\Big)\Big(1 - \underbrace{p(0)h}_{\to 0} - \underbrace{p''(0)\frac{h^3}{24}}_{\to 0} + \underbrace{\bar{\bar o}(h^3)}_{\to 0}\Big)$$

$$\le \frac{1}{nh}\,p(0)(1+\bar{\bar o}(1)), \quad h\to 0$$

$$\mathrm{MSE}(\hat p_n(0)) = \mathrm{Var}(\hat p_n(0)) + \mathrm{Bias}^2(\hat p_n(0)) = \Big(\frac{p(0)}{nh} + \frac{h^4(p''(0))^2}{576}\Big)(1+\bar{\bar o}(1)) \quad h\to 0$$

(ii) Является ли $\hat p_n(0)$ несмещённой оценкой $p(0)$? Является ли $\hat p_n(0)$ состоятельной оценкой $p(0)$?

$$\mathrm{Bias}(\hat p_n(0)) = \frac{p''(0)h^2}{24} + \bar{\bar o}(h^2) \neq 0 \Rightarrow \text{смещ., но асимпт. несм. при } h \to$$

$$\mathrm{Var}(\hat p_n(0)) = \frac{p(0)}{nh}(1+\bar{\bar o}(1)) \longrightarrow 0 \Leftrightarrow \underline{nh\to\infty} \Rightarrow \text{сост. при } nh\to\infty$$

$$\mathbb{E}\big[(\hat p_n(0) - \mathbb{E}[\hat p_n(0)])^2\big] \to 0$$

Задача 2

T2 Допустим, что дана выборка из нормального распределения с нулевым средним и дисперсией σ^2. Вычислите значение параметра bandwidth, минимизирующее AMISE (asymptotic mean integrated squared error) ядерной оценки плотности, построенной на основе гауссовского ядра

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

$$\text{KDE}: \hat p_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x-x_i}{h}\Big)$$

$$\mathrm{AMISE}(\hat p_n(x)) = \frac{h^4}{4}\int_{\mathbb{R}}(p''(x))^2dx\Big(\int_{\mathbb{R}}x^2K(x)dx\Big)^4 + \frac{1}{nh}\int_{\mathbb{R}}K^2(x)dx$$

$$\frac{\partial\,\mathrm{AMISE}(\hat p_n(x))}{\partial h} = 0$$

$$h_{opt} = \frac{1}{n^{1/5}}\left(\frac{\int_{\mathbb{R}}K^2(x)dx}{\int_{\mathbb{R}}(p''(x))^2dx\,\big(\int_{\mathbb{R}}x^2K(x)dx\big)^4}\right)^{1/5}$$

• $\int_{\mathbb R} K^2(x)\,dx = \int_{\mathbb R} \frac{1}{2\pi} e^{-x^2}\,dx = \frac{1}{2\pi}\int_{\mathbb R} e^{-x^2}\,dx = \frac{1}{2\sqrt{\pi}}$

• $\left(\int_{\mathbb R} x^2 K(x)\,dx\right)^4 = \left(\int_{\mathbb R} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx\right) = \left[\xi \sim N(0,1)\right] = \left(\mathbb E\xi^2\right)^4 = 1^4 = 1$

• $p''(x) = \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\right)'' = \left(-\frac{x}{\sqrt{2\pi}\sigma^3} e^{-\frac{x^2}{2\sigma^2}}\right)' = -\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma^3} + \frac{x^2}{\sqrt{2\pi}\sigma^5} e^{-\frac{x^2}{2\sigma^2}}$

$\Rightarrow (p''(x))^2 = \frac{e^{-x^2/\sigma^2}}{2\pi\sigma^6} - 2\frac{x^2}{2\pi\sigma^8} e^{-x^2/\sigma^2} + \frac{x^4}{2\pi\sigma^{10}} e^{-x^2/\sigma^2}$

$\int_{\mathbb R} (p''(x))^2\,dx = \int_{\mathbb R} \frac{e^{-x^2/\sigma^2}}{2\pi\sigma^6}\,dx - 2\int_{\mathbb R} \frac{x^2}{2\pi\sigma^8} e^{-x^2/\sigma^2}\,dx + \int_{\mathbb R} \frac{x^4}{2\pi\sigma^{10}} e^{-x^2/\sigma^2}\,dx$

$= \frac{1}{2\sqrt{\pi}\sigma^5}\underbrace{\int_{\mathbb R} \frac{1}{\sqrt{2\pi}\cdot\frac{\sigma}{\sqrt2}} e^{-\frac{x^2}{2\cdot\sigma^2/2}}\,dx}_{\int_{\mathbb R}\varphi_{N(0,\sigma^2/2)} = 1} - \frac{2}{2\sqrt{\pi}\sigma^7}\underbrace{\int_{\mathbb R} \frac{x^2}{\sqrt{2\pi}\cdot\frac{\sigma}{\sqrt2}} e^{-\frac{x^2}{2\cdot\sigma^2/2}}}_{\mathbb E[X^2],\ X\sim N(0,\sigma^2/2)} + \frac{1}{2\sqrt{\pi}\sigma^9}\underbrace{\int_{\mathbb R} \frac{x^4}{\sqrt{2\pi}\cdot\frac{\sigma}{\sqrt2}} e^{-\frac{x^2}{2\cdot\sigma^2/2}}\,dx}_{\mathbb E X^4,\ \ldots}$

$X \overset{d}{=} \frac{\sigma}{\sqrt2}\xi \Rightarrow \mathbb E[X^2] = \frac{\sigma^2}{2}\mathbb E\xi^2 = \frac{\sigma^2}{2}$

$= \frac{\sigma^4}{4}\mathbb E\xi^4 = \frac{3\sigma^4}{4}$

$= \cancel{\frac{1}{2\sqrt{\pi}\sigma^5}} - \cancel{\frac{1}{2\sqrt{\pi}\sigma^5}} + \frac{3}{8\sqrt{\pi}\sigma^5} = \frac{3}{8\sqrt{\pi}\sigma^5}$

$\Rightarrow h_{opt} = \frac{1}{n^{1/5}}\left(\frac{8\sqrt{\pi}\sigma^5}{2\sqrt{\pi}\cdot 3}\right)^{1/5} = \frac{\sigma}{n^{1/5}}\underline{\left(\frac{4}{3}\right)^{1/5}}$ → "nrd" in R

Задача 3

T3 Пусть $\psi_0, \psi_1, \ldots$ - ортонормированный базис Лежандра в пространстве $L^2([-1,1])$. Докажите, что функция

$$K(x) = \sum_{m=0}^{n} \psi_m(0)\psi_m(x)\mathbb{I}\{|x| \le 1\}$$

является ядром порядка n в том смысле, что $\int_{\mathbb R} K(x)dx = 1$ и $\int_{\mathbb R} K(x)x^k dx = 0$ для $k = 1, \ldots, n$. → $\int_{\mathbb R} x^k K(x)\,dx,\ k \in [\![0, n]\!]$

$\int_{-1}^{1} \psi_i(x)\psi_j(x) = \begin{cases} 0, & i \ne j \\ 1, & i = j \end{cases}$

k → ψ_k имеет степень k

$x^k \in L^2([-1,1]),\quad x^k = \sum_{i=0}^{k} \psi_i(x)\,\alpha_{ik}$

$\int_{\mathbb R} x^k K(x)\,dx = \int_{-1}^{1} \sum_{i=0}^{k} \psi_i(x)\alpha_{ik} \sum_{m=0}^{n} \psi_m(0)\psi_m(x)\,dx = \sum_{i=0}^{k}\sum_{m=0}^{n} \alpha_{ik}\psi_m(0) \overbrace{\int_{-1}^{1} \psi_i(x)\psi_m(x)\,dx}^{=0,\ i\ne m}$

$= \underbrace{\overbrace{\sum_{i=0}^{\min\{k,n\}}}^{k} \alpha_{ik}\psi_m(0)}_{x^k|_{x=0}} \cdot \overbrace{1}^{\int_{-1}^{1}\psi_m^2(x)dx} = \begin{cases} 0, & k \in [\![1, n]\!] \\ 1, & k = 0 \end{cases}$

Задача 4

T4 Оценка кросс-валидации параметра h определяется как точка минимума функции

$$\hat{J}(h) = \int \hat{p}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{p}_{(-i)}(x_i),$$

где $\hat{p}_{(-i)}$ - оценка, построенная по всем значениям, кроме i−го (leave-one-out estimate). Для ядерной оценкой плотности $\hat{p}_n$, представьте функцию $\hat{J}(h)$ в виде функции, зависящей от выборки $x_1, .., x_n$ только через разности $x_i - x_j$, $i, j = 1, \dots, n$.

$$\int_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx = \int_{\mathbb{R}} (\hat{p}_n(x))^2 dx - 2\underbrace{\int_{\mathbb{R}} \hat{p}_n(x)\, p(x)\, dx}_{= \mathbb{E}[\hat{p}_n(X)],\ X \sim p(x)} + \cancel{\int_{\mathbb{R}} p^2(x) dx} \to \min$$

$$\mathbb{E}[\hat{p}_n(X)] - \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{(-i)}(x_i)$$

$$\hat{J}(h) = \int_{\mathbb{R}} \frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\right)^2 dx - \underbrace{\frac{2}{n}\sum_{i=1}^{n}\sum_{j \neq i} \frac{1}{(n-1)h} K\left(\frac{x_i - x_j}{h}\right)}$$

$$= \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{(n-1)h} K\left(\frac{x_i - x_j}{h}\right) - \frac{2}{\cancel{n}} \cdot \cancel{n} \cdot \frac{1}{(n-1)h} K(0)$$

$$\bullet\ \frac{1}{n^2h^2}\int_{\mathbb{R}} \sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{x - x_j}{h}\right) dx = \left[\begin{array}{l} u = \frac{x - x_i}{h} \\ x = uh + x_i \end{array}\right] = \frac{1}{n^2h^2}\int_{\mathbb{R}} \sum_{i=1}^{n}\sum_{j=1}^{n} K(u)\, K\left(u + \frac{x_i - x_j}{h}\right) h\, du$$

$K(u) = K(-u)$

$$\overset{K\text{-чёт}}{=} \frac{1}{n^2h^{\cancel{2}}}\sum_{i=1}^{n}\sum_{j=1}^{n} \int_{\mathbb{R}} K(u)\, K\left(\frac{x_j - x_i}{h} - u\right) \cancel{h}\, du$$

$$f * g(x) = \int_{\mathbb{R}} f(x-u)\, g(u)\, du \Big\} = \frac{1}{n^2h}\sum_{i=1}^{n}\sum_{j=1}^{n} K * K\left(\frac{x_j - x_i}{h}\right)$$

$$\hat{J}(h) = \frac{1}{n^2h}\sum_{i,j=1}^{n} K * K\left(\frac{x_j - x_i}{h}\right) - \frac{2}{n}\sum_{i,j=1}^{n} \frac{1}{(n-1)h} K\left(\frac{x_i - x_j}{h}\right) + \frac{2}{(n-1)h} K(0)$$

$$(n-1) \updownarrow n$$

$$h = \underset{h>0}{\operatorname{argmin}}\, \hat{J}(h) \to \text{"ucv"}$$

$$\text{"bcv"}: \underset{h>0}{\operatorname{argmin}} \left\{ \frac{1}{n^2h}\sum_{i,j=1}^{n}\left(K * K\left(\frac{x_j - x_i}{h}\right) - 2K\left(\frac{x_j - x_i}{h}\right)\right) + \frac{2}{(n-1)h} K(0) \right\}$$